Incident Report

Sharedband Service outage 16th, 28th and 29th March 2017

Updated 31st March 2017

Executive Summary

On the 16th March, Sharedband experienced a critical service outage which impacted on services hosted in the London data centre.

At 13:08UTC, Sharedband's primary upstream internet link was disconnected. In an event of this nature, traffic should re-route via other internet links that Sharedband have available. This did not happen and a large portion of Sharedband's services were isolated.

A ticket with the provider was opened and the provider acknowledged that they were working to resolve an outage on their network. The incident was resolved by the provider at 13:51UTC.

At 17:18UTC on the 28th March, the BGP connection between Sharedband and the Provider was reset. The connection was only down for 10 seconds, however the impact to service was 20 minutes due to slow routing convergence on the Providers network.

At 03:15UTC there was another disruption to service due to a maintenance event on the Providers network that was not communicated to Sharedband. Disruption to service was approximately 20 minutes.

Sharedband have been working with the provider to understand the events that caused the loss of traffic to our network during these events.

On the evening of 30th March, Sharedband had a telephonic meeting with the Global Director of Support of the Provider. In this meeting, the findings of the Providers investigation were presented to Sharedband.

The nature of the maintenance on the 16th March was the installation of new core network equipment. This new equipment was to increase their capacity at the London PoP. The new equipment has encountered a bug in the processing of routing protocol updates. Whenever there is a break in connectivity to the Provider, their equipment CPU reaches 100% utilisation and routing protocol convergence suffers extended delays during this time. They have seen convergence take up to an hour. This directly impacts traffic to Sharedband every time there is an upset in connectivity in the Providers network.

The Provider has opened a P1 support ticket with their routing equipment provider and investigations from the vendor are continuing as to the cause of the CPU utilisation and slow routing reconvergence.

Sharedband will be in constant communication with the Provider to prepare for any future maintenance where there may be disruption. Sharedband will manually withdraw route announcements to the Provider before their maintenance so that traffic to Sharedband's services can reroute quicker via alternative connections.

Table of Contents

Incident Details

16th March

On the 16th March at 13:08UTC, Sharedband Infrastructure engineers were alerted to an outage by monitoring systems. Upon investigation, it was found that the primary Internet uplink was in a down state. The upstream provider was contacted via telephone and a ticket opened. At 13:30UTC the provider acknowledged that they had lost power to a switch and were in the process of restoring power.

At 13:44 our Providers Support Department confirmed that power had been restored to the switch. They also confirmed that they could see that our connection was still down and that their engineers were trying to restore the link.

At 13:51UTC the link was restored and traffic was restored to Sharedband's services.

28th March

On March the 28th at 17:18UTC, Sharedband's BGP session to the Provider was reset and re-established within 10 seconds. However traffic to Sharedband did not reroute as expected as with the outage on the 16th March. Traffic took 20 Minutes to re-establish and by 17:38 service was restored.

29th March

On the 29th March at 03:15, traffic to Sharedband's network was disrupted due to a maintenance event on the Providers network. Traffic levels returned within 20 minutes of the initial outage.

Root Cause Analysis

16th March

Since the outage, Sharedband have been awaiting a Root Causes Analysis report from the Provider to better understand the events that happened before and during the incident. The RCA was provided to us on the 25th March.

The provider had a Maintenance window between 02:00 and 06:00UTC on Thursday morning. During this maintenance window, they were replacing core routing equipment in their London Point-of-Presence. The maintenance was completed successfully within the communicated change window. However, engineers continued to work in the cabinet where the equipment was located. The nature of this work was the installation of new Power Distribution Units. Engineers mistook the status of a power indicating light and thought there was power to the PDU. They connected the secondary power supply of the switch Sharedband was connected to, to the new PDU.

At 13:08, the primary power cable to the switch was pulled in order to move it to a new PDU but because the secondary power supply was not active, the switch lost power. Once power was restored, Sharedband's network port did not initialise properly when the switch restored and network engineers had to disable and re-enable the port for the connection to Sharedband to re-establish again.

Impact to Sharedband was 40 minutes due to slow network reconvergence on the provider's network.

28th March

At 17:18, the provider briefly lost connectivity between their Core router and the switch that Sharedband connect to. The Provider are investigating the possibility that unplanned work at the hosting facility was taking place and a fibre cable to their cabinet was briefly disrupted, causing the loss of connectivity between their router and switch and ultimately causing the slow routing reconvergence to happen again.

29th March

At 03:15UTC the provider was busy with planned maintenance that was not meant to be customer affecting and therefore was not communicated to Sharedband. In this maintenance, additional connectivity was added between the Provider and the London Internet Exchange. This caused the same slow routing reconvergence to happen.

Mitigation

After meeting with the Provider on the 30th March the Provider has provided Sharedband with additional information about the outages and the mitigation put in place to prevent these incidents from happening again.

We now know that the root cause of the slow routing protocol reconvergence is due to a software bug on the new core routing equipment installed in London on the morning of the 16th March. A case has been opened with the equipment vendor and they are investigating the Provider's configurations as well as the software for a cause of the high CPU utilisation when BGP routing changes happen.

The Provider also admits that the planned maintenance on the morning of the 29th should not have gone ahead and there was a lapse in communication to the engineer performing the maintenance as he was on an inbound flight to London to perform the maintenance and he never received the communication not to proceed.

The Provider has admitted that implementation of the new equipment was rushed and changes went ahead that should not have. They have committed to a complete review of their internal processes and communication to us around any future changes to ensure that future occurrences of the past few weeks events to not reoccur.

Sharedband will also proactively move traffic away from the provider before any maintenance windows to prevent the slow routing convergence seen since the introduction of the new equipment. This will be in effect until such time as we have confirmation that the cause of the high CPU utilisation and slow BGP reconvergence has been addressed.